Method of evaluating image-recovery algorithms based on task performance

K. M. Hanson

Los Alamos National Laboratory, MS P940, Los Alamos, New Mexico 87545

Received October 25, 1989; accepted February 21, 1990

A method of evaluating image-recovery algorithms is presented that is based on the numerical computation of how well a specified visual task can be performed on the basis of the reconstructed images. A Monte Carlo technique is used to simulate the complete imaging process including generation of scenes appropriate to the desired application, subsequent data taking, image recovery, and performance of the stated task based on the final image. The pseudo-random-simulation process permits one to assess the response of an image-recovery algorithm to many different realizations of the same type of scene. The usefulness of this method is demonstrated through a study of the algebraic reconstruction technique (ART), a tomographic reconstruction algorithm that reconstructs images from their projections. The task chosen for this study is the detection of disks of known size and position. Task performance is rated on the basis of the detectability index derived from the area under the receiver operating characteristic curve. In the imaging situations explored, the use of the nonnegativity constraint in the ART dramatically increases the detectability of objects in some instances, particularly when the data consist of a limited number of noiseless projections. Conversely, the nonnegativity constraint does not improve detectability when the data are complete but noisy.

INTRODUCTION

In every indirect imaging application it is necessary to choose an image-recovery algorithm to obtain a final image. This choice becomes critically important when the available data are limited and/or are noisy. Several classes of measures on which to base image-recovery algorithms have been employed in the past.[1] There are those based on how close the reconstructed image is to the original one, such as the conventional measure of minimum rms difference between the reconstruction and the original image. Experience teaches us that this does not always seem to be correlated with the usefulness of images and so does not help one to choose an algorithm. There are measures based on how closely the estimated reconstruction reproduces the input data. The most popular of these measures, based on least-squares residual (or minimum chi squared), is known often to be ill conditioned or, even worse, ill posed.[1] To make the problem more tractable, it is often proposed to constrain the least-squares objective in some way. Further, there are measures that combine the foregoing measures, such as maximum *a posteriori* reconstruction, which balances the agreement with the data against the relationship of the reconstruction to the known ensemble probability distributions.[2]

The fundamental tenet adopted in this paper is that the overall purpose of the imaging procedure is to provide certain specific information about the object or scene under investigation. Consequently, in the approach to algorithm evaluation presented here, an algorithm is to be judged on the basis of how well one can perform stated visual tasks, using the reconstructed images.

The method presented can help one to answer the perennial question asked of tomographers: How many projections are needed? The proper response to such a question is the following: needed to do what? The answer depends on the type of scene that one is dealing with, the magnitude of the noise in the data, and the kind of information that one desires from the reconstruction, to wit, the visual task to be performed.

For linear imaging systems the effects of image noise on task performance can be predicted for a variety of tasks, as, for example, treated in Ref. 3. The masking effects of measurement noise are truly random in nature. The random-noise process makes each set of measurements different, even when the scene being imaged does not change. Contrarily, the effects of artifacts on task performance are not so easily predicted. Some kinds of artifact appear as fixed patterns and do not often behave like stationary noise. However, those created by an insufficient number of measurements can manifest themselves as seemingly unpredictable irregularities that look like noise, but, in a strict sense, they are not. These patterns are determined by the scene being imaged. Therefore it is necessary to deal with realistic scenes to test how well an algorithm dispenses with artifacts. For example, the objects in the scene are typically randomly placed relative to the discretely sampled measurements as well as to the reconstruction grid. Both of these positionings can affect the reconstruction. Thus a single realization of a simple scene is completely inadequate for judging a reconstruction algorithm. It is necessary to obtain a statistically meaningful average of the response of an algorithm to many realizations of the ensemble of scenes with which it must cope. It is unclear whether such a global approach to task performance is amenable to theoretical treatment. The implied averaging over discrete samplings is difficult to handle analytically although some results can be derived.[4] It is *not* properly taken into account by the assumption of an effective modulation transfer function to characterize sampling, as is so often employed. Furthermore, it would be difficult to deal with nonlinear reconstruction or task performance algorithms. To overcome these deficiencies, the proposed method is based on computer simulation of scenes

0740-3232/90/071294-11$02.00

appropriate to the desired application, the subsequent data taking, and the analysis of the data. A Monte Carlo technique, one employing pseudorandom numbers to generate its results, is used in this simulation process because it can readily provide the above-noted variations within the ensemble. Furthermore, any new source of uncertainty can easily be incorporated into the simulation by simply selecting the appropriate variable by means of a pseudorandom number.

METHOD OF EVALUATING TASK PERFORMANCE

The proposed method of evaluating image-recovery algorithms[5] employs a Monte Carlo technique[6] to simulate the entire imaging process from the beginning to the final task performance. To begin, one randomly generates representative scenes and the corresponding sets of measurements. The specified tasks are then performed, using the reconstructed scenes. Finally, the accuracy of the task performance is evaluated. The advantage of this numerical approach is that it readily handles complex imaging situations, nonstationary imaging characteristics, and nonlinear reconstruction algorithms. Its major disadvantage is that it provides an evaluation that is valid only for the specific imaging situation investigation.

Figure 1 shows the basic steps of the proposed method of evaluating the task performance based on an ensemble of images. The proposed method proceeds as follows. First, the whole problem must be completely specified:

(a) Define the class of scenes to be imaged, including as much complexity as exists in the intended application. Variations in scene from one realization to another should be fully specified.

(b) Define the geometry of the measurements. The deficiencies in the measurements such as blur, uncertainties in the geometry, and uncertainties in the measurements (noise) should be specified. Variations of these uncertainties with position, as well as intercorrelations between them, could be included.

(c) Define clearly the task that is to be performed. The task might be simple detection of a known object against a known background, for example. Alternatively, it could be discrimination between two types of object, or something more complex, such as multiple discrimination or parameter estimation. The fundamental assumptions in effect must be explicitly stated.

(d) Define the method of task performance. This should be consistent with the intended application. If the task is to be performed by computer, then the intended analysis algorithm may be used. If the task is to be performed by a human observer, some approximation to the way in which a human interprets an image should be used. Alternatively, a maximum-likelihood algorithm (ideal observer) may be employed to define the best possible performance (under the prevailing assumptions made about the extent of auxillary information).

The simulation procedure is then performed by doing the following:

(e) Create a representative scene and the corresponding measurement data by means of a Monte Carlo simulation technique. All variations in scene content and uncertainties in the measurements are included by means of pseudorandom selection of the uncertain and variable parameters.

(f) Reconstruct the scene with the algorithm being tested.

(g) Perform the specified task, using the reconstructed image.

(h) Repeat steps (e)–(g) a sufficient number of times to obtain the necessary statistics on the accuracy of the task performance.

Finally, determine how well the task has been performed, on the average:

(i) Evaluate the task performance. For binary discrimination tasks (of the yes or no variety), a receiver operating characteristic (ROC) curve[7] may be generated. The proper measure of how well the task is performed should be based on what is important in the intended application. In a precise treatment, one might use Bayes's method to estimate the total risk, incorporating the relative costs of making false or true conclusions.[7,8] For parameter estimation tasks, the standard measure of rms error might be appropriate.

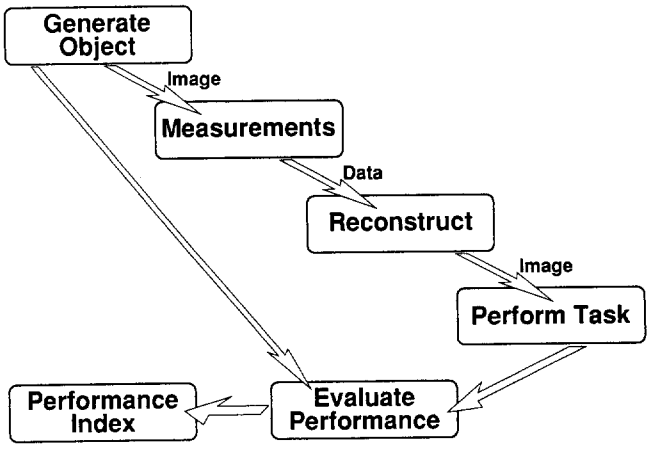

Fig. 1. Diagram of the Monte Carlo procedure employed to evaluate numerically the performance of a visual task.

ALGEBRAIC RECONSTRUCTION TECHNIQUE

The algebraic reconstruction technique[9] (ART) is an iterative algorithm that reconstructs a function from its projections. It has proved to be a successful tomographic reconstruction algorithm, particularly when there is a limited number of projections available. Assume that N individual projection measurements are made of the unknown function $\mathbf{f}$, which is considered a vector. The ith measurement is written as

$$\mathbf{g}_i = \mathbf{H}_i\mathbf{f}, \qquad i = 1, \ldots, N, \tag{1}$$

where $\mathbf{H}_i$ is the corresponding row of the measurement matrix. It should be realized that the originally suggested choices of 0 and 1 for the elements of $\mathbf{H}$ is unwise.[10] The precise weights used are not intrinsic to the algorithm. In modern practice the elements of $\mathbf{H}$ are calculated as line or

strip integrals of the reconstruction by interpolating between the discretely sampled grid points.[11,12]

The ART algorithm proceeds as follows. An initial guess is made, for example, $\mathbf{f}^0 = 0$. Then the estimate is updated by iterating on the individual measurements taken in turn:

$$\mathbf{f}^{k+1} = \mathbf{f}^k + \lambda^k \mathbf{H}_i^T \left(\frac{\mathbf{g}_i - \mathbf{H}_i \mathbf{f}^k}{\mathbf{H}_i^T \mathbf{H}_i} \right), \tag{2}$$

where $\mathbf{f}^k$ is the kth estimate of the image vector $\mathbf{f}$, $i = k \bmod(N) + 1$, and λ^k is a relaxation factor for the kth update. In constrained ART algorithms a nonnegativity constraint is enforced by setting any component of $\mathbf{f}^{k+1}$ to zero that has been made negative by the above updating procedure. We use the index K to indicate the iteration number [$K = \text{int}(k/N)$], which in the standard nomenclature corresponds to one pass through all N measurements. We express the relaxation factor as

$$\lambda^K = \lambda_0 (r_\lambda)^{K-1}. \tag{3}$$

There is little guidance on the choice of the relaxation factor in the literature. A value of unity is often suggested and used. In the absence of constraints, such a choice forces the reconstruction to agree with the last measurement used for updating. It is known[13] that if a unique solution to the measurement equations exists, the ART algorithm converges to it in the limit of an infinite number of iterations, provided that $2 > \lambda > 0$. If many solutions exist, the ART algorithm converges to the one with minimum norm. Censor *et al.*[14] have shown that the unconstrained ART algorithm ultimately converges to a minimum-norm least-squares solution, which is desirable for inconsistent (noisy) data, if the relaxation factor approaches zero slowly enough. In this research we will investigate the ART algorithm, employing 10 iterations. Although 10 iterations are not enough to ensure complete convergence to the solution, this choice is in the range of the number of iterations often employed with the ART algorithm.[9] We will use 1.0 and 0.8 as nominal values for λ_0 and r_λ for problems involving a limited number of views or projection sets and 0.2 and 0.8 for problems involving many (~100) views. These choices are fairly representative for unconstrained ART reconstructions and will suffice for the present demonstration of the proposed method of evaluation. It is possible to use this evaluation method to find the best choice of relaxation parameters for a specific problem.[15]

APPLICATION TO THE EVALUATION OF THE NONNEGATIVITY CONSTRAINT

The usefulness of the nonnegativity constraint in the ART algorithm will now be explored to demonstrate how the proposed method can be used. It should be noted that such a constraint makes the response of the reconstruction algorithm nonlinear. As such, the task performance in the presence of either noise or artifacts is not amenable to linear analysis. By way of introduction to the choice of properties for the scenes to be studied here, Fig. 2 shows how the simple streak artifacts that arise in the tomographic reconstruction of a single disk from a limited number of projections superimpose to form a complex background pattern when several objects are present. In one sense the seemingly random fluctuations in the background are not truly noise because they would be exactly reproduced if another set of projections was obtained with the objects in the same place. However, the pattern changes with the positions of the objects. So, in another sense, the artifacts are stochastic if the objects are randomly placed in the scene. This simple observation points to the need to consider many realizations of the kind of scene in order to adequately evaluate task performance in the presence of artifacts.

For the present example, each scene is assumed to consist

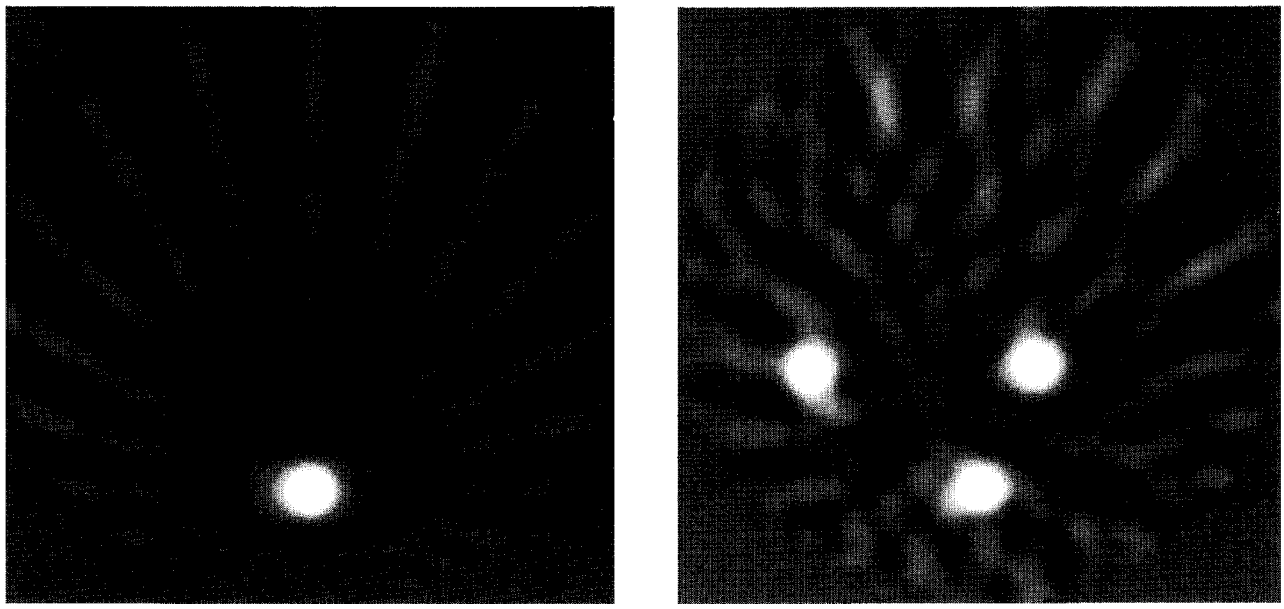

Fig. 2. Reconstructions of one (left) and three (right) dots from 12 noiseless parallel projections covering 180 deg obtained with the unconstrained ART algorithm. The incoherent sum of simple streak artifacts from three dots produces noise-like fluctuations in the background.

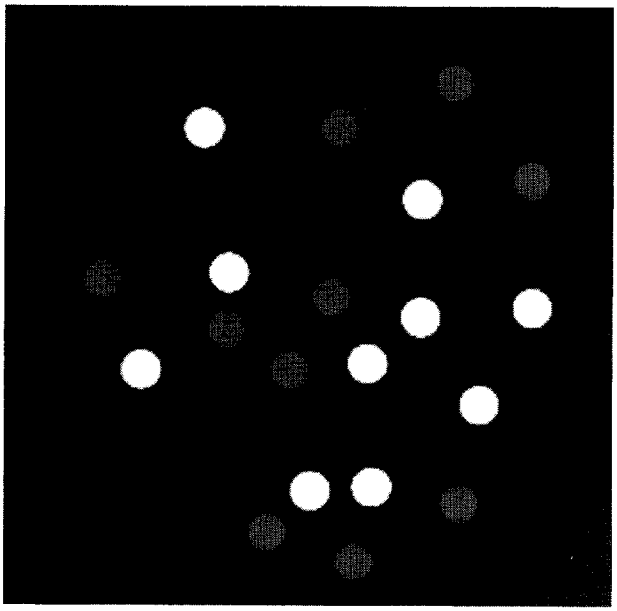

Fig. 3. The first randomly generated scene consisting of 10 high-contrast and 10 low-contrast disks. The present estimation of detectability is based on an average over 10 similar scenes.

of a number of nonoverlapping disks placed on a zero background. To maximize the problems caused by artifacts in this study, we assume that each scene contains 10 high-contrast disks of amplitude 1.0 and 10 low-contrast disks with amplitude 0.1. The task to be performed is the detection of the low-contrast disks. In limited data-taking circumstances, the high-contrast disks produce serious artifacts in the reconstructions, making it difficult to detect the low-contrast ones. The disks are randomly placed within the circle of reconstruction, which has a diameter of 128 pixels in the reconstructed image. The diameter of each disk is 8 pixels. To guarantee that the disks do not touch one another in the reconstruction, a 3-pixel-wide buffer region surrounds each disk. The first of the series of images generated for these tests is shown in Fig. 3. In this computed tomographic problem, the measurements are assumed to consist of a specified number of parallel projections, each containing 128 samples. The above choice for the kind of scene to be studied provides a situation in which the nonnegativity constraint is likely to have a substantial effect. In some of the test cases described below, random noise is added to the projection measurements. For these, a Gaussian-distributed random number generator with zero mean is used. This means that negative values for the projections are possible, even though the object itself is nonnegative. While this may seem absurd to theoreticians, it is not at variance with many experimental situations. For example, in transmission tomography in which the projections are measured through the attenuation of x rays, the path length is derived from the ratio of a measured x-ray intensity to that expected for no object. The measured intensity values will vary, at least because of counting statistics, about less than the expected intensity, yielding path lengths that fluctuate about and below zero.

The results of reconstructing Fig. 3 from 12 noiseless views spanning 180 deg by using 10 iterations of the ART algorithm are shown in Fig. 4. The seemingly random fluctuations in the background are artifacts produced by the limited number of projections. At first sight, it appears that the nonnegativity constraint improves the reconstruction considerably by reducing the confusion caused by the fluctuations in the background. However, some of the low-contrast disks have not been reproduced. Also, there remain many fluctuations in the background that may mislead one to suspect the presence of disks where none exist in reality. Thus, on the basis of this single reconstructed scene, one

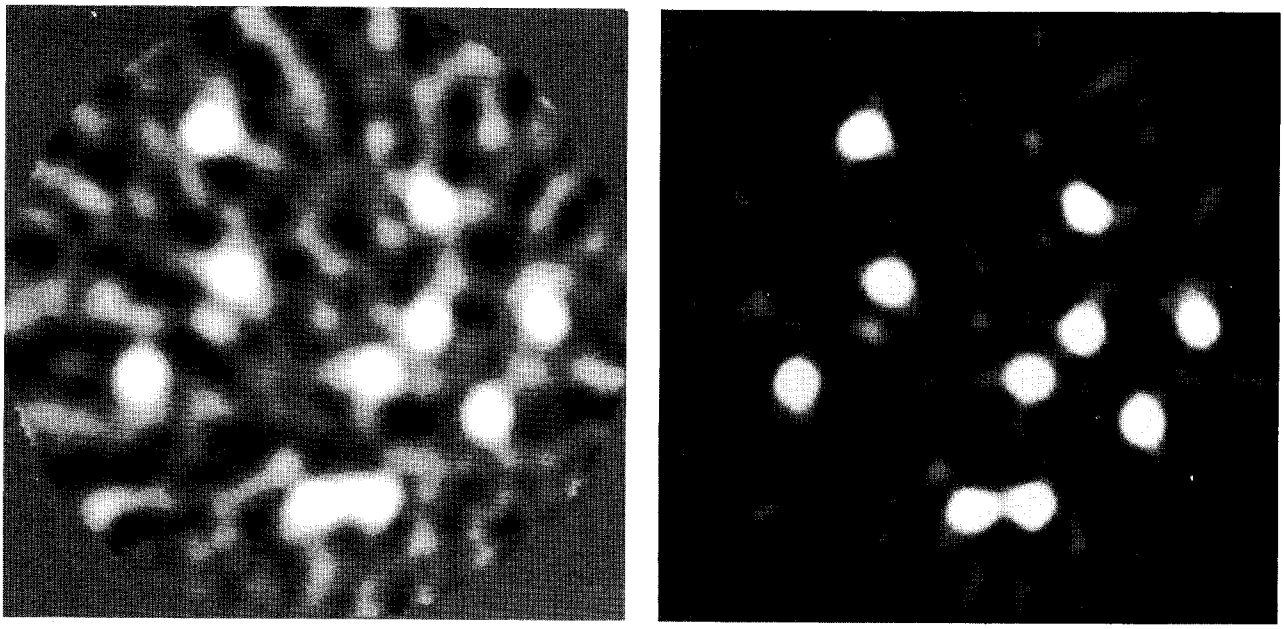

Fig. 4. Reconstructions of Fig. 3 from 12 noiseless parallel projections subtending 180 deg obtained with 10 iterations of the ART algorithm with (right) and without (left) the nonnegativity constraint. These images are displayed with high contrast to reveal the low-contrast disks of interest.

cannot say with certainty whether the nonnegativity constraint improves the detection of the low-contrast disks. The question is basically a statistical one. A statistically significant comparison between reconstructions with and without the constraint must be made to assess its value.

We now present some basic elements of signal-detection theory[7] to facilitate the analysis of the binary discrimination task, namely, is the disk present or not? To make the detection task as simple as possible, it will be assumed that the position of a possible disk and the background is completely known beforehand. The first step is to define a scalar decision variable, which is to be used to make the decision. The likelihood ratio between the two alternatives yields an optimum decision variable. When the image is corrupted by additive uncorrelated Gaussian noise, the optimum decision variable is the inner product between the expected signal shape and the actual data, which is, of course, identical to using a matched filter.[7] Then the detection task is performed by stating that a disk is present when the value of the decision variable is above a chosen threshold value. Consider the frequency distribution of the decision variables obtained at locations at which the objects of interest are known to be present. The probability that the presence of a disk is correctly detected, called the true-positive probability, is estimated by the area that is under this frequency distribution and above the threshold. The probability of falsely stating that a disk is present, the false-positive probability, is the area that is above the threshold and under the corresponding frequency distribution for locations at which no object exists. As the threshold is lowered, both the true-positive rate and the false-positive rate increase. The resulting variation sweeps out the ROC curve,[7] which is a plot of the true-positive probability versus the false-positive probability. The ROC curve completely summarizes binary discrimination task performance. According to Bayes, an optimum value of the threshold value, one that minimizes the overall risk or cost, can be chosen on the basis of the relative costs associated with correctly and incorrectly detecting disks. When one is dealing with human observers, these frequency distributions are not explicitly observable; the choice of the threshold is implicitly made by the observer. The access to the frequency distributions afforded by the present computational approach is advantageous because these distributions represent the fundamental data. The full ROC curve is easily generated once the frequency distributions are calculated.

To perform the stated task of detection in the present study, we assumed that the average of the reconstruction over the area of the disk is an appropriate decision variable. This average is a good approximation to the optimal matched filter. However, it ignores the blurring effects of the finite resolution of the discretely sampled reconstruction. It also does not take into account the known correlation in the noise in computed tomographic reconstructions[16] that have been derived from projections containing uncorrelated noise. Nor does it take into account the effects of the nonnegativity constraint on the characteristics of the noise. After reconstruction, the average is calculated over each region in which a low-contrast disk is known to exist. The result of doing this for the reconstructions of 10 different scenes, each containing 10 low-contrast disks and approximately 30 separate disk regions that are taken from the background, is displayed as a frequency graph in this decision variable ψ in Fig. 5(a). The graph for the averages over each region for which no disk exists is plotted as well.

Figure 6 shows the ROC curve generated directly from the frequency graphs in Fig. 5. Comparison between the ROC curves produced by the unconstrained ART algorithm and by the constrained ART algorithm shows that the nonnegativity constraint has dramatically enhanced the performance of this detection task. We will base our summary measure of the detectability on the area under the ROC curve A, which is known to be the same as the fraction of correct scores that would be obtained in a two-alternative forced-choice experiment.[17] As is characteristic of all summary measures, however, it ignores the details of the shape of the ROC curve, which might be important if different costs were involved in making false-positive responses than in making false-negative ones. The area is estimated here by applying the trapezoidal rule to the ROC curve generated with medium to finely binned histograms. The areas under the ROC curves in Fig. 6 are 0.738 with no constraint compared with 0.930 with the constraint. The area under the ROC curve may be expressed in terms of an effective index for detectability d_A:

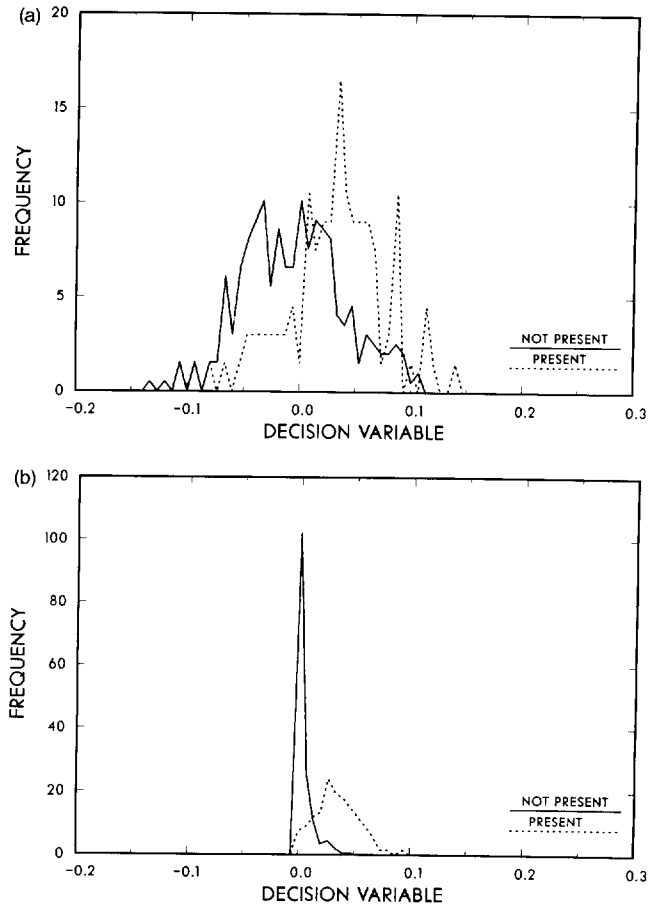

Fig. 5. Frequency graphs of the decision variable (the average over a circular region) evaluated at positions where a low-contrast disk is known to exist (dashed curve) and where none exists (solid curve) for ART reconstructions (a) without the nonnegativity constraint and (b) with the constraint. These results summarize the performance obtained from reconstructions from 12 views for 10 randomly generated scenes.

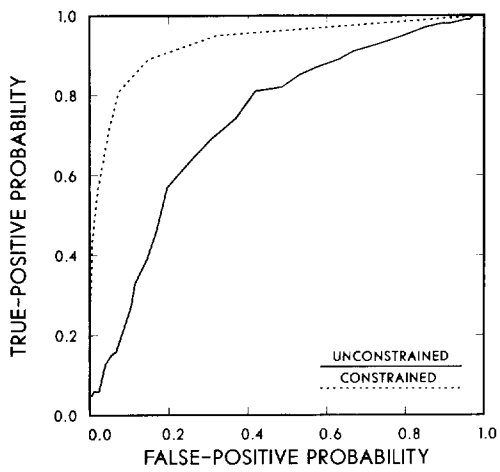

Fig. 6. ROC curves derived from the frequency graphs that are shown in Fig. 5 for unconstrained ART (solid curve) and for constrained reconstructions (dashed curve). The nonnegativity constraint is seen to improve detectability markedly, as its ROC curve is always significantly above that for no constraints.

$$d_A = 2 \operatorname{erf} c^{-1}[2(1 - A)], \tag{4}$$

where erf c^{-1} is the inverse of the complement of the error function erf $c(x) = 1 - (2/\sqrt{\pi})\int_0^x \exp(-t^2)dt$.[18,19] In Ref. 19 this index is designated by $z(A)$ and the area under the ROC curve by $P(A)$. For Fig. 6 a value for d_A of 0.901 is obtained for the reconstructions without the constraint and 2.092 with the constraint. Thus the use of the nonnegativity constraint has increased the detectability by 132%, in this case of a limited number of views. The nonnegativity constraint decidedly improves this measure of detectability.

An alternative index of detectability can be derived directly from the frequency distributions. For one to be able to distinguish between the ensemble of low-contrast disks and the background, it is clearly desirable for these two frequency distributions to be separated as much as possible. The degree of separation between the two distributions is often characterized by the detectability index d' (called d_a in Ref. 19), given by

$$d' = \frac{\bar{\psi}_1 - \bar{\psi}_0}{\left(\frac{\sigma_1^2 + \sigma_0^2}{2}\right)^{1/2}}, \tag{5}$$

where $\bar{\psi}_1$ and σ_1 are the mean and the rms deviation of the frequency distribution, respectively, when the object is present and those with subscript 0 are the values when the object is not present. This index is sometimes called the detection signal-to-noise ratio. Equation (5) is normalized to be the same as d_A for Gaussian-shaped frequency distributions. For the frequency graphs that are shown in Fig. 5(a) corresponding to the unconstrained reconstructions, d' is 0.871. The corresponding frequency graphs for constrained reconstructions are presented in Fig. 5(b). It is seen that, owing to the nonnegativity constraint, the frequency graph for the background regions piles up against zero. Both frequency graphs are narrower than those for the unconstrained reconstructions. In fact, the rms widths of the two frequency graphs are quite different. The detectability index d' for the constrained reconstructions is 2.054. Care should be exercised in using this measure of detectability,[19] as it is equivalent to d_A only when the two underlying frequency distributions are Gaussian. So if the stated task is binary discrimination between the disks and the background and the area under the ROC curve is deemed to be the appropriate performance index, then d' should not be employed in place of d_A without verification of the Gaussian shape of the distributions in the decision variable. Despite the rather non-Gaussian distributions observed in Fig. 5(b), d' is close to d_A. Note that the average reconstruction value over a disk provides an estimate of the amplitude of the disk. The relative accuracy of such an estimate is $\sigma_1/\bar{\psi}_1$. Thus there is an intimate connection between the detection task considered here and the task of amplitude estimation.[3,20]

The relative accuracy of the two indices of detection is worth mentioning. The statistical accuracy of results obtained by the Monte Carlo method must always be considered because these results are calculated by averaging over a finite number of discrete occurrences called events. Of the two measures of detectability presented above, d' has better statistical accuracy, as it is calculated by using all the events in the two frequency distributions. By the usual method of propagating errors the rms uncertainty in d' is easily estimated to be approximately

$$\sigma_{d'} = \left[\frac{2 + (d'/2)^2}{n}\right]^{1/2}, \tag{6}$$

where n is the number of events in each graph, which is assumed here to be the same for both. The only events that contribute to the ROC curve are those that lie in the range of the decision variable that is common to both frequency graphs that are to be distinguished. Thus the calculation of A, and therefore of d_A, is based on only a fraction of the n events. The rms uncertainty in d_A given by Simpson and Fitter[19] is

$$\sigma_{d_A} = \sqrt{4\pi}\left[\frac{A(1 - A)}{n}\right]^{1/2} \exp\left[\left(\frac{d_A}{2}\right)^2\right]. \tag{7}$$

The accuracy of d_A is only slightly worse than that of d' for small d'. But as d' grows, the accuracy of d_A soon becomes much worse because the number of events in the region common to both frequency distributions diminishes eventually to zero, as the frequency distributions ultimately become completely disjoint. Even though it is not so accurate as d', d_A is the relevant measure for the simple binary detection task, also known as the signal-known-exactly detection task. Clearly for this task the only range of ψ for which there is any confusion in detection is the region common to the two frequency distributions. What happens outside that region is unimportant for this particular task.

The above estimates are those of the rms deviation in the results that would be observed for many repeated realizations of the same imaging situation. However, they may not properly gauge the significance of the change in detectability index that might be observed when two different algorithms are compared on the basis of exactly the same data sets. It is observed that a high degree of correlation often exists between the reconstructions obtained with different algorithms when one starts from exactly the same data. This correlation is advantageous because it increases the significance of an intercomparison made between two algorithms for a given number of trials. However, the significance of

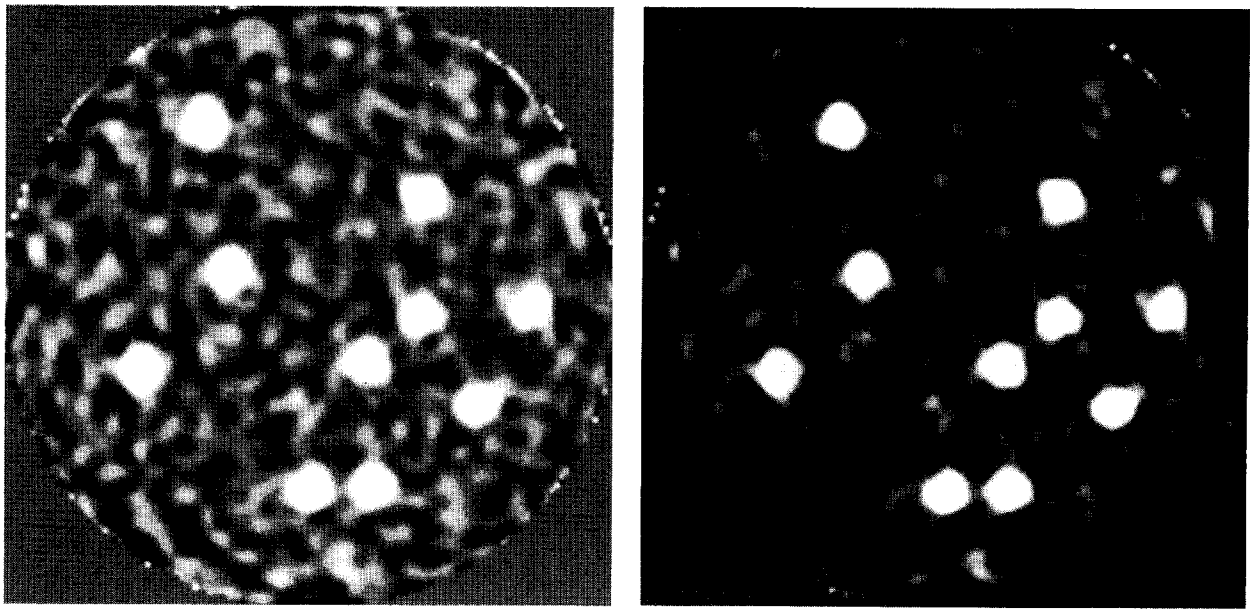

Fig. 7. Reconstructions of Fig. 3 from 100 noisy parallel projections subtending 180 deg obtained with 10 iterations of the ART algorithm with (right) and without (left) the nonnegativity constraint. The added noise is Gaussian distributed with a rms deviation of 8.

the difference may be determined only by a detailed statistical analysis of the data as, for example, provided by the ROC analysis code CORROC developed by Metz[21] or by the technique suggested by Hanley and McNeil.[22]

Figure 7 shows the reconstructions obtained for an essentially complete set of data, 100 projections covering 180 deg but with noise added to the projection data. The rms value of the noise is 8. For comparison, the peak projection value of a low-contrast disk is 0.80. Again it appears that the nonnegativity constraint produces a visible improvement in the reconstructions by almost completely eliminating the noise in the background. However, a disk-by-disk visual comparison between the constrained reconstruction and the known original scene (Fig. 3) indicates that the ability to detect each disk is questionable. There is clearly a need to accumulate statistics on many objects and scenes to determine whether the constraint has improved detectability. The frequency graphs in Fig. 8 provide the desired statistical summary. For the unconstrained reconstructions, the frequency graphs for both regions appear to be Gaussian shaped with essentially identical rms widths, as expected, since the unconstrained reconstruction algorithm is a stationary linear process. The characteristics of the frequency graphs in Fig. 8 are much the same as those discussed above. The detectability index d' obtained from the frequency graphs for the unconstrained reconstructions is 1.995, and d_A derived from the corresponding ROC curve is 1.964. The detectability in this situation can be calculated on the basis of the rms noise in the projection data under the assumption that the background is known.[3] The resulting d' is approximately 2.3. This is slightly larger than the value obtained above, which one expects because no account is taken of the correlations intrinsic to noise in computed tomographic reconstructions.[16] Thus, in this case of complete and noisy data, the ART algorithm achieves nearly full statistical efficiency, as is expected of the filtered backprojection algorithm.[23]

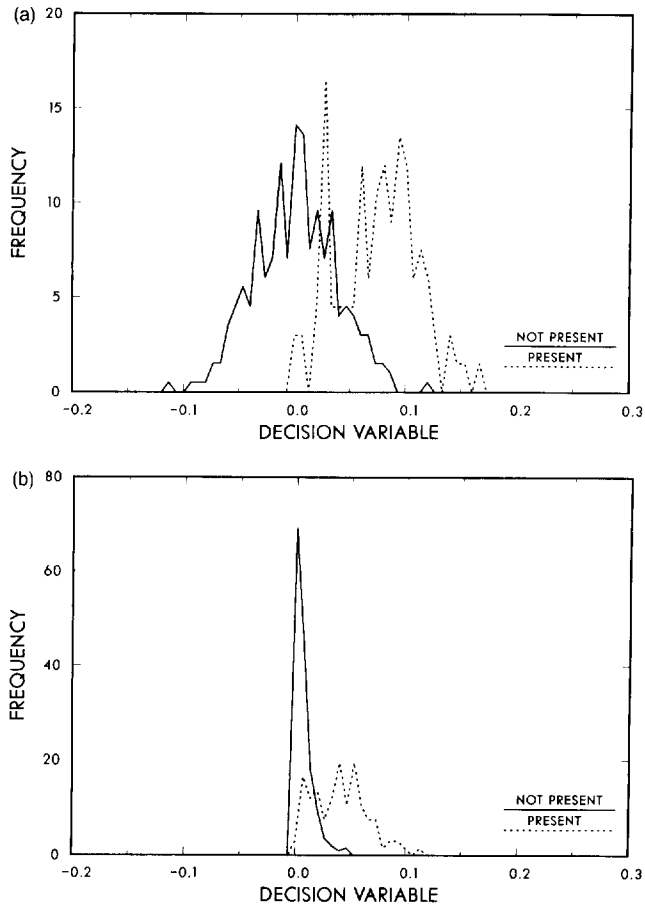

Fig. 8. Frequency graphs of the decision variable (the average over a circular region) evaluated where a low-contrast disk is known to exist (dashed curve) and where none exists (solid curve) for ART reconstructions (a) without the nonnegativity constraint and (b) with the constraint. These results summarize the performance obtained from reconstructions of 10 randomly generated scenes from 100 views with an rms noise value of 8.

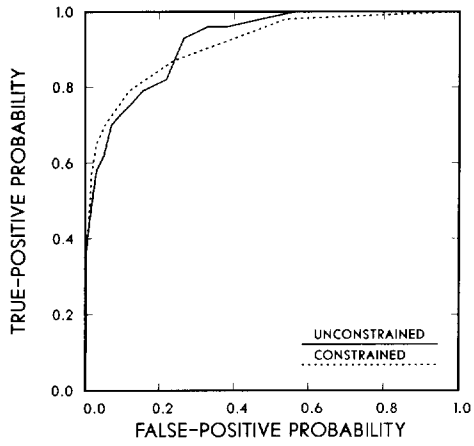

Fig. 9. The ROC curves derived from the frequency graphs that are shown in Fig. 8 for unconstrained ART (solid curve) and for constrained reconstructions (dashed curve). The nonnegativity constraint does not alter detectability in a statistically significant way.

The nonnegativity constraint has a profound effect on the frequency graphs in Fig. 8. Negative values for the decision variable are not possible, so the frequency graph for the background region becomes spiked at zero. The mean value of the decision variable of the frequency graph for the low-contrast disk regions is reduced by almost a factor of 2. However, the ROC curves (Fig. 9) for the unconstrained and constrained cases are essentially identical, indicating that there is no change in discrimination between the background and the low-contrast disks. The implication is that, under an appropriate transformation of the decision variable, the two frequency graphs are essentially unchanged in their region of commonality, despite the obvious changes in gross graph shapes. From the frequency graphs of the constrained reconstructions, d' is 1.825 and d_A derived from the area under the ROC curve is 1.985. The statistical uncertainty in these values for d' is approximately 0.141 by the more general form of Eq. (6) for an unequal number of events in the two frequency graphs, and the statistical uncertainty in d_A is approximately 0.206. The difference between these two indices of detectability is almost statistically significant.

Summarizing the above discussion, the d_A values for the ROC curves are 1.964 and 1.985, without and with the constraint, respectively, each with an estimated statistical uncertainty of 0.206. Thus we might conclude that the nonnegativity constraint increases the detectability by no more than 25%, based on a 1.65 standard-deviation limit for a confidence level of 5%. However we expect the statistical accuracy in the comparison between these two values to be much better than this because of the strong correlation between the unconstrained and the constrained reconstructions mentioned above. It is found that the frequency graphs of the differences of the decision variables for the two reconstruction sets, taken region by region, have rms widths that are almost half those for the individual frequency graphs. One concludes that the nonnegativity constraint does not increase the detectability by more than approximately 13%. This conclusion is perhaps contrary to what we might be led to believe on the basis of the improved visual appearance of Fig. 7(b). As this result is counterintuitive, it deserves closer investigation in future research.

Table 1 tabulates the detectability indices obtained under various data-taking conditions. The estimated accuracies of d_A take into account the fact that the number of events in the frequency graphs are 100 and nearly 300. The nonnegativity constraint is seen to be generally useful. The constraint is particularly helpful when the data are limited by deficiencies in the measurement geometry. It has little effect when the data are complete but noisy. The CPU time required to calculate the entries in the table look as long as 1 h on a VAX 8700, which is about four times faster than a VAX 785.

Table 2 summarizes other results obtained for the same conditions. As noted above, d' has better statistical accuracy than d_A, ranging from 0.114 to 0.196 for the entries in the table, assuming 100 and 300 events in the two frequency graphs. The rms error is the rms difference between the 10 reconstructions and their corresponding original images.

Table 1. Summary of the Effect of the Nonnegativity Constraint on the Detectability Index d_A Determined from the Area under the ROC Curve for Various Kinds of Projection Data[a]

No. Projections	$\Delta\theta$ (Deg)	RMS Noise	d_A Without Constraint	d_A With Constraint	Relative Improvement in d_A
100	180	8	1.964 ± 0.205	1.985 ± 0.206	+1%
100	180	4	4.113 ± 0.826	4.514 ± 1.223	+10%
8	180	0	0.458 ± 0.145	0.732 ± 0.149	+60%
12	180	0	0.901 ± 0.153	2.092 ± 0.215	+132%
16	180	0	1.937 ± 0.202	5.322 ± 3.02	+175%
16	90	0	1.176 ± 0.161	2.319 ± 0.238	+97%
32	90	0	1.244 ± 0.164	3.329 ± 0.434	+168%
16	180	2	1.626 ± 0.182	2.616 ± 0.279	+16%
16	180	1	1.845 ± 0.199	3.965 ± 0.722	+115%
16	90	1	1.265 ± 0.163	2.165 ± 0.221	+71%

[a] These results were obtained from tests made on the computed tomographic reconstructions of 10 randomly generated scenes for various kinds of deficiency in the data. In all cases 10 iterations of the ART algorithm were used with nominal relaxation factors, as explained in the text.

Table 2. Summary of the Effect of the Nonnegativity Constraint on Other Measures of Reconstruction Quality for the Same Situations Tabulated in Table 1

No. Projections	$\Delta\theta$ (Deg)	RMS Noise	d' Without Constraint	d' With Constraint	RMS Error Without Constraint	RMS Error With Constraint	L1 Error Without Constraint	L1 Error With Constraint	RMS Residual Without Constraint	RMS Residual With Constraint
100	180	8	1.995	1.825	0.101	0.063	0.068	0.020	3.55	3.86
100	180	4	4.001	4.032	0.069	0.056	0.040	0.017	1.79	1.94
8	180	0	0.464	0.653	0.125	0.095	0.068	0.033	0.064	0.763
12	180	0	0.871	2.054	0.109	0.074	0.063	0.024	0.067	0.553
16	180	0	1.960	4.782	0.092	0.061	0.053	0.018	0.071	0.362
16	90	0	1.122	2.050	0.116	0.093	0.057	0.029	0.188	0.466
32	90	0	1.184	3.227	0.112	0.082	0.055	0.025	0.161	0.331
16	180	2	1.653	2.372	0.098	0.065	0.059	0.020	0.503	0.876
16	180	1	1.860	3.698	0.093	0.062	0.055	0.019	0.259	0.537
16	90	1	1.105	1.795	0.117	0.094	0.055	0.030	0.351	0.619

The L1 error is similar but is calculated as the average of the absolute value of the differences. It is concluded from this table that the nonnegativity constraint enhances detectability when the data are incomplete. An equally fundamental result is that the detectability for complete, but noisy, data is not improved by the nonnegativity constraint. An opposite conclusion would be drawn from the rms and the L1 norm errors, both of which indicate that the constrained reconstructions are significantly closer to the original images. These measures probably fail to be indicative of task performance because the fluctuations in the constrained reconstructions are no longer position invariant, that is, stationary. Another reason to distrust such summary measures of reconstruction fidelity is that they do not distinguish between different spatial frequencies. As such they cannot be indicative of the performance of specific tasks.[3]

It is noted that d' is not much different from d_A in most of the situations tested, even though the frequency distributions for the constrained reconstructions never possess Gaussian shapes. This is encouraging because d' has better statistical accuracy than d_A, particularly for large d', and is more likely to be a continuous function of the parameters that can be varied in the reconstruction procedure. Thus d' is a desirable performance index for the purpose of optimizing a reconstruction technique, if one bears in mind the caveats, stated above, concerning the connection between d' and d_A. If d_A is deemed the appropriate performance index, good relative accuracy in its estimation can be easily obtained only when it falls in the range from approximately 1 to 3. Thus the design of the imaging situation must be carefully adjusted to keep d_A within that range. Although the relative accuracy of d' keeps getting better as d' gets larger, at some point its relevance might be questioned, because discrimination becomes virtually certain. Furthermore, systematic effects of the simulation procedure become more relevant as the statistical errors decrease.

DISCUSSION

We have presented a method to test the effectiveness of reconstruction algorithms. This method is based on a Monte Carlo simulation of the complete imaging process from the composition of the original scene to the final interpretation of the reconstructed image. The goal of the simulation is to estimate the accuracy with which a specified task can be performed on the basis of the reconstructions. This method accords with the notion that an algorithm can be properly evaluated only by trying it out on a statistically meaningful sample of trials. A major benefit of the Monte Carlo technique is that new effects may be easily added. On the other hand, only the overall effect of all the conditions is observed. It may be difficult to determine the relative contributions of individual effects. The Monte Carlo simulation technique is particularly useful in situations that do not lend themselves to an analytic approach. It can provide a good statistical sampling over all the uncontrollable variables in the problem. An example is the typical case of the effect of discrete sampling on signal analysis, as in the problem of the detection of small objects. In this example it is desirable to average the detectability over all possible positions of the object relative to the discrete measurements and the reconstruction grid.[4] The Monte Carlo method is perfect for this.

We have seen that the nonnegativity constraint is often beneficial for the specific problem addressed here—detection of low-contrast disks in computed tomographic reconstructions. This constraint is particularly helpful when the data consist of a limited number of noiseless projections. However, when the data are complete but degraded by additive noise, the nonnegativity constraint does not improve detectability. Some improvement is attained in intermediate circumstances when the data are both incomplete and noisy. One can abstract these results by concluding that the use of prior knowledge (that the image must be nonnegative) improves the usefulness of reconstructions containing artifacts created by the null space associated with a lack of measurements.[2] On the other hand, when the defects in the reconstruction are a consequence of noise in the measurements, the nonnegativity constraint is of no help. We hypothesize that the value of the nonnegativity constraint will generally depend on which of these characterizes the deficiency in the data. In previous research, we found that the effectiveness of the nonnegativity constraint can be significantly enhanced by choosing the relaxation parameters used in the ART algorithm to optimize the detectability.[15]

It is possible to obtain misleading results by assuming too simple a task. For example, by consideration of the binary–

singlet discrimination (Rayleigh) task applied directly to the acquired quantum-limited data, Wagner *et al.*[24] came to the puzzling conclusion that a large square aperture is preferable to a coded aperture. A subsequent study[25] corroborated and extended the original analysis. The rub is the simultaneous assumption of only a single object, which can have just two configurations, and a known background. Under these assumptions, binary discrimination can logically be made by comparing the raw data against the two alternative signal shapes. This task can be performed without the necessity of reconstructing the scene. It has been shown[26] that the presence of an unknown, slowly varying background reduces both the detectability and the Rayleigh discriminability for large apertures substantially more than for apertures that are approximately the size of the object. The smaller aperture is preferred because it reduces the mixture of unknown background with the signal. The latter analysis was also performed solely on the basis of the measurement data. It is conjectured that coded apertures will prove to provide better performance than the simple square aperture as more variability and corresponding uncertainty is introduced into the problem. Examples of such increased complexity for the Rayleigh task are unknown position and orientation of the binary object and the presence of other unknown objects in the scene. With the Monte Carlo evaluation method, one can easily accommodate such additional complications by numerically estimating task performance in many reconstructed images. It is anticipated that the presence of artifacts in the reconstructions will reveal a major hole in the "grand gaping aperture" argument.[25]

From the above we see that a fundamental distinction exists between performing binary discrimination tasks on the basis of directly measured data as opposed to using reconstructed images. An analysis based on the raw data amounts to a calculation of the propagation of random errors for the particular measurement matrix. This analysis places an upper limit on the accuracy of binary discrimination. This upper limit can be attained in situations in which a complete set of data is available, for which the nasty null space and, hence, artifacts in the reconstructions are eliminated. Logically speaking, it is also applicable in the rare situations in which the signal and the background are completely known *a priori*, as assumed in the above-mentioned studies. In cases involving noiseless data, such an analysis always implies perfect discriminability. In the present study we have observed the contrary; detectability based on reconstructions from noiseless data can be far from perfect because of the artifacts produced by the (unknown) collection of objects in the scene in conjunction with a limited number of projections. The appearance of these artifacts is equivalent to the lack of knowledge of the background in the projection data for such a problem. Although the reconstruction procedure is supposed to separate the objects in the reconstructed scene, it can do so only if enough projection data are available.

As the detection task specified in the present example is truly simple and not closely related to many real problems, it would be worthwhile to explore more complex and interesting tasks.[20,27,28] Clearly, the definition of the task and the method used to perform it are extremely important for a reliable conclusion. But it may be difficult to define precisely many interesting real-life tasks. For example, how does one approach the problem of detection of a lesion in a radiograph? Another aspect of real imaging systems is that they almost invariably must handle multiple types of tasks in many types of images. The solution in terms of the Monte Carlo method is to invoke a performance index that takes a weighted average over as many different types of tasks and images as necessary to produce a relevant measure of efficacy.

The use of a nonnegativity constraint leads to a bias in reconstructed images. Thus there is probably a need to acknowledge a lack of information about the background surrounding an object. Inclusion of an unknown background leads to a weighting function composed of a positive central region surrounded by a negative annulus, a so-called center–surround mask. An alternative way to handle an unknown background is to employ a more general, least-squares fitting approach in which the reconstructed image data are fitted to an assumed object signal plus a constant or slowly varying background.[28] The fitting approach can be used to estimate many other object parameters, for example, its position.[20]

A worthwhile extension of the present research would be to pursue alternative choices for the decision variables for the purpose of improving performance. For example, one might consider a weighted average of the reconstruction values over a local region, much the same as the simple circular weight function used here but with considerably more flexibility. The optimal weights might be determined by using half the simulated reconstructions as a training set and the second half to estimate the task performance index. It would probably be too difficult to handle completely general weight assignments, but, with suitable restrictions on the number of variables used to specify the weights, it might be feasible. The optimal choice of decision variable might depend on the reconstruction procedure. If this line of research were pursued, it would be reasonable to compare the performance of one algorithm against another only on the basis of the best-decision procedure that could be achieved with each. If a human observer is to be the final interpreter of the images, the method used to perform the visual task must be correctly modeled to mimic the human observer, which might not be so easy to accomplish numerically.[27] It is interesting to note that the human observer probably cannot take into account a known background value in an absolute way. Thus it does not make sense to build this prior knowledge into the task. In principle, it is of course feasible to incorporate a human observer directly into the present method, that is, to entreat a human interpreter to make the required decision on the basis of a sequence of reconstructed images that have been generated by the computer. However, it should be realized that reliable results can be achieved only through careful preparation and painstaking training of the observers.[29–31] The latter aspect of dealing with human observers must be taken seriously because artifacts keep changing character as the algorithm changes.

Clearly, this approach of random simulation is generally applicable to testing and evaluating any or all aspects of the entire imaging chain from scene generation to the final method of task performance. Possibly a fruitful line of research that can be addressed by using this approach is the optimization of the imaging system, either in terms of its

individual parts or in its entirety. If many parameters are to be varied in the optimization, one must be concerned about the stability of the optimization procedure. Regularization may be required in order to stabilize the search for the optimum. For example, the optimization function could be augmented by a sum of squares of the deviations of the parameters from some standard values.

ACKNOWLEDGMENTS

I have profitted greatly from discussions with many of my colleagues, including Harrison Barrett, Arthur Burgess, Allen Mathews, Charles Metz, Kyle Myers, Robert Wagner, and Rollin Whitman. In particular, Kyle Myers and Robert Wagner have kept me out of trouble on numerous occasions. This research was supported by the U.S. Department of Energy under contract W-7405-ENG-36 and by Thomson CGR (now GE-CGR).

REFERENCES

1. H. C. Andrews and B. R. Hunt, *Digital Image Restoration* (Prentice-Hall, Englewood Cliffs, N.J., 1977).
2. K. M. Hanson, "Bayesian and related methods in image reconstruction from incomplete data," in *Image Recovery: Theory and Application*, H. Stark, ed. (Academic, Orlando, Fla., 1987).
3. K. M. Hanson, "Variations in task and the ideal observer," in *Application of Optical Instrumentation in Medicine XI*, G. D. Fullerton, ed., Proc. Soc. Photo-Opt. Instrum. Eng. **419,** 60–67 (1983).
4. K. M. Hanson, "The detective quantum efficiency of CT reconstruction: the detection of small objects," in *Application of Optical Instrumentation in Medicine VII*, G. D. Gray, ed., Proc. Soc. Photo-Opt. Instrum. Eng. **173,** 291–298 (1979).
5. K. M. Hanson, "Method to evaluate image-recovery algorithms based on task performance," in *Medical Imaging II*, R. H. Schneider and S. J. Dwyer, eds., Proc. Soc. Photo-Opt. Instrum. Eng. **914,** 336–343 (1988).
6. R. Y. Rubinstein, *Simulation and the Monte Carlo Method* (Wiley, New York, 1981).
7. A. D. Whalen, *Detection of Signals* (Academic, New York, 1971).
8. I. J. Good, *Good Thinking—The Foundations of Probability and its Applications* (U. Minnesota Press, Minneapolis, Minn., 1983).
9. R. Gordon, R. Bender, and G. Herman, "Algebraic reconstruction techniques for three-dimensional electron microscopy and x-ray photography," J. Theor. Biol. **29,** 471–481 (1970).
10. P. Gilbert, "Iterative methods for the three-dimensional reconstruction of an object from projections," J. Theor. Biol. **36,** 105–117 (1972).
11. G. T. Herman and A. Lent, "Iterative reconstruction algorithms," Comput. Biol. Med. **6,** 273–294 (1976).
12. K. M. Hanson and G. W. Wecksung, "Local basis-function approach to computed tomography," J. Opt. Soc. Am. **24,** 4028–4039 (1985).
13. G. T. Herman, A. Lent, and P. H. Lutz, "Relaxation methods for image reconstruction, Commun. ACM **21,** 152–158 (1978).
14. Y. Censor, P. P. B. Eggermont, and D. Gordon, "Strong underrelaxation in Kaczmarz's method for inconsistent systems," Numer. Math. **41,** 83–92 (1983).
15. K. M. Hanson, "POPART—performance optimized algebraic reconstruction technique," in *Visual Communications and Image Processing '88: Third in a Series*, T. R. Hsing, ed., Proc. Soc. Photo-Opt. Instrum. Eng. **1001,** 318–325 (1988).
16. K. M. Hanson, "Detectability in computed tomographic images," Med. Phys. **6,** 441–451 (1979).
17. D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics* (Kreiger, Huntington, N.Y., 1966).
18. J. A. Swets, "ROC analysis applied to the evaluation of medical imaging techniques," Invest. Radiol. **14,** 109–121 (1979).
19. A. J. Simpson and M. J. Fitter, "What is the best index of detectability?" Psychol. Bull. **80,** 481–488 (1973).
20. K. M. Hanson, "Optimization for object localization of the constrained algebraic reconstruction technique," in *Medical Imaging III: Image Formation*, S. J. Dwyer, R. G. Jost, and R. H. Schneider, eds., Proc. Soc. Photo-Opt. Instrum. Eng. **1090,** 146–153 (1989).
21. C. E. Metz, "Statistical analysis of ROC data in evaluating diagnostic performance," in *Multiple Regression Analysis: Applications in the Health Sciences*, D. Herbert and R. Myers, eds. (American Institute of Physics, New York, 1986), pp. 365–384.
22. J. A. Hanley and B. J. McNeil, "A method of comparing the areas under receiver operating characteristic curves derived from the same cases," Radiology **148,** 839–843 (1983).
23. K. M. Hanson, "On the optimality of the filtered backprojection algorithm," J. Comput. Assist Tomogr. **4,** 361–363 (1980).
24. R. F. Wagner, D. G. Brown, and C. E. Metz, "On the multiplex advantage of coded source/aperture photon imaging," in *Digital Radiography*, W. R. Brody, ed., Proc. Soc. Photo-Opt. Instrum. Eng. **314,** 72–76 (1981).
25. K. J. Myers, R. F. Wagner, D. G. Brown, and H. H. Barrett, "Efficient utilization of aperture and detector by optimal coding, in *Medical Imaging III: Image Formation*, S. J. Dwyer, R. G. Jost, and R. H. Schneider, eds., Proc. Soc. Photo-Opt. Instrum. Eng. **1090,** 164–175 (1989).
26. H. H. Barrett, J. P. Rolland, K. J. Myers, and R. F. Wagner, "Detection and discrimination of known signals in inhomogeneous, random backgrounds," in *Medical Imaging III: Image Formation*, S. J. Dwyer, R. G. Jost, and R. H. Schneider, eds., Proc. Soc. Photo-Opt. Instrum. Eng. **1090,** 176–182 (1989).
27. R. F. Wagner, K. J. Myers, D. G. Brown, M. J. Tapiovaara, and A. E. Burgess, "Higher-order tasks: human vs. machine performance, in *Medical Imaging III: Image Formation*, S. J. Dwyer, R. G. Jost, and R. H. Schneider, eds., Proc. Soc. Photo-Opt. Instrum. Eng. **1090,** 183–194 (1989).
28. K. M. Hanson, "Optimization of the constrained algebraic reconstruction technique for a variety of visual tasks, in *Proceedings of Information Processing in Medical Imaging XI*, D. A. Ortendahl and J. Llacer, eds. (Liss, New York, to be published).
29. A. E. Burgess and H. Ghandeharian, "Visual signal detection. I. Ability to use phase information," J. Opt. Soc. Am. A **1,** 900–905 (1984).
30. A. E. Burgess and H. Ghandeharian, "Visual signal detection. II. Signal-location identification," J. Opt. Soc. Am. A **1,** 906–910 (1984).
31. A. E. Burgess, "Visual signal detection. III. On Bayesian use of prior information and cross correlation," J. Opt. Soc. Am. A **2,** 1498–1507 (1985).